Emotion Prediction Based on Facial Expressions and Speech.

Author(s): -

Kavita Virk (KV28@myscc.ca)

Kunal Ghaware (KG88@myscc.ca)

Harsh Trivedi (HT35@myscc.ca)

Pranav Jadhav (PJ25@myscc.ca)

Under the guidance of:

Prof. Umair Durrani

CogniXR health

Content

ABSTRACT

The aim of this project is to build a machine learning model which can predict the emotions of a person based on the facial expression and speech of that person and then deploying this model on the webpage of CogniXR health. CogniXR health is start-up which provides telehealth services. In this project we have build 2 separate models which can predict the emotion of the person. One model will used to predict the emotions from the image while the other one will be used to predict the emotion from the speech of the person. The source of images would be the video file uploaded by the therapist. The API will capture the video and defragment into number of images and audio sets which will be used as input to our models. Finally, the API will deploy our models on the webpage.

INTRODUCTION

In ancient times people used to visit doctors but there were no digital devices or tools used by doctors at those times. In today's modern world, there are many devices such as stethoscope, ECG machine, endoscope, etc., which are used by the doctors to diagnose and monitor patients' health. But all these devices are used to monitor physical health of a patient. What about the mental health of the patient? Is there any tool to monitor the mental health of a patient which could be used by the therapist to identify the emotional status of a patient and help the therapist to cure the patient? That's exactly where we came into action. Under the guidance of CogniXR health and our professor Prof. Umair Durrani we tried to find a solution for this problem.

CogniXR health, a start up company is an all-in-one telehealth platform designed for accelerating the practice of therapists' and improving standards of care. This platform is built for patients' engagement, used by EAP employee assistance programs, schools, and professionals in intervention in private, public, and Non-Profit organizations.

In this project, CogniXR health expected from us to build a model for them which could recognise the emotions of the client based on their facial expressions and their speech and then, deploy this model on CogniXR health's website.

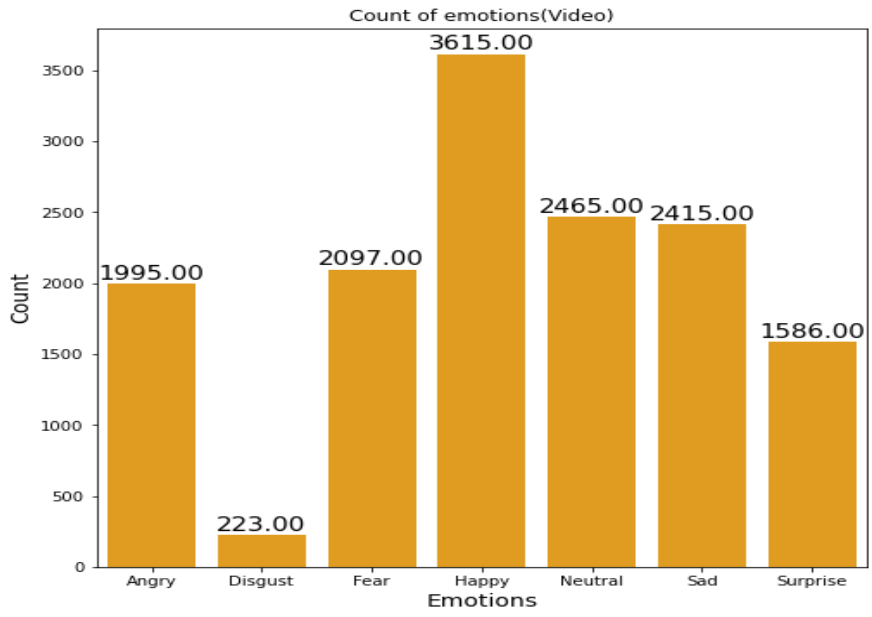
We already had alpha model from the previous group of students who worked the same project. Alpha model means a model which is been already created by someone but not yet tested on user's platform. But there were some limitations in that model. That model was biased to some extent. So, we decided to either improve that model or build a new model.

Hence, we developed two different models. First model, the image classifier model is built to classify the images into various emotions and the second model, the audio classifier model is built to identify and categories the emotions based on the audio input to the model. This is done using python.

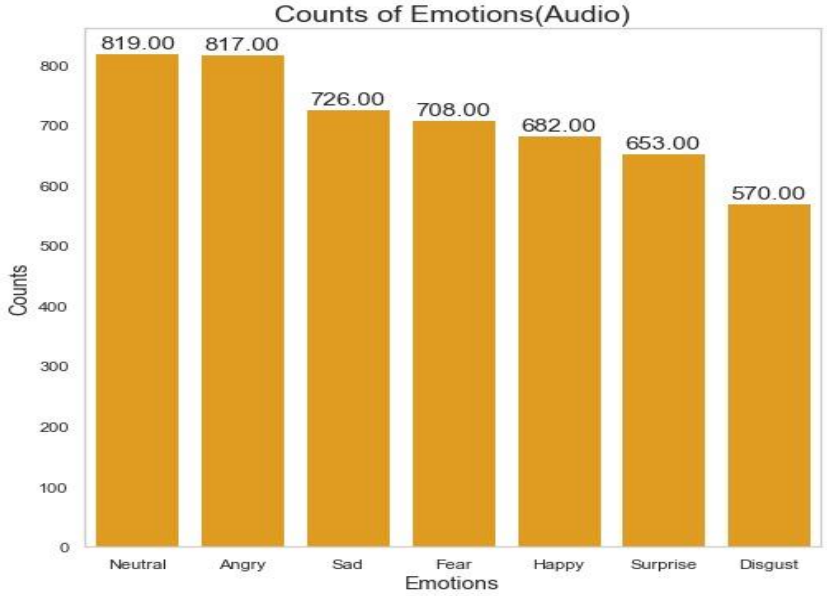
After developing both the model, an API is used to deploy those models on CogniXR health's website. This is build using python.

Description of dataset

There are 2 datasets used in this project. One for the emotion recognition from image and other for the emotion recognition from the audio.



The dataset used for the image recognition is downloaded from the Kaggle website. The name of the dataset is FER-2013. There are total 35887 images in this dataset which is broadly divided into training and testing dataset. The training dataset has 28,709 images while the testing dataset has 7,178 number of images. However, we have used only 50% of the training data for training our model due to computing constraints. These datasets are further divided into 7 different categories as follows: Angry (13.86%), Disgust (1.55%), Fear (14.57%), Happy (25.11%), Neutral (17.12%), Sad (16.78%), Surprise (11.02%). All the images are in grayscale of size 48 X 48 pixels.



The datasets used for Audio emotion recognition is downloaded from Kaggle website. To make the dataset multilingual, we have mixed different datasets with specific languages together. There are total 5 different languages used for our project and that are as follows: German, Italic, Urdu, Arabic and English. All the audio files are spoken by both genders. The files are in WAV format. There are total 4975 Audio files in our dataset from which 20% of data we used or testing. This dataset is further divided into 7 different classes which are as follows: Angry (16.42%), Disgust (11.45%), Fear (14.23%), Happy (13.70%, Neutral (16.46%), Sad (14.59%), Surprise (13.12%).

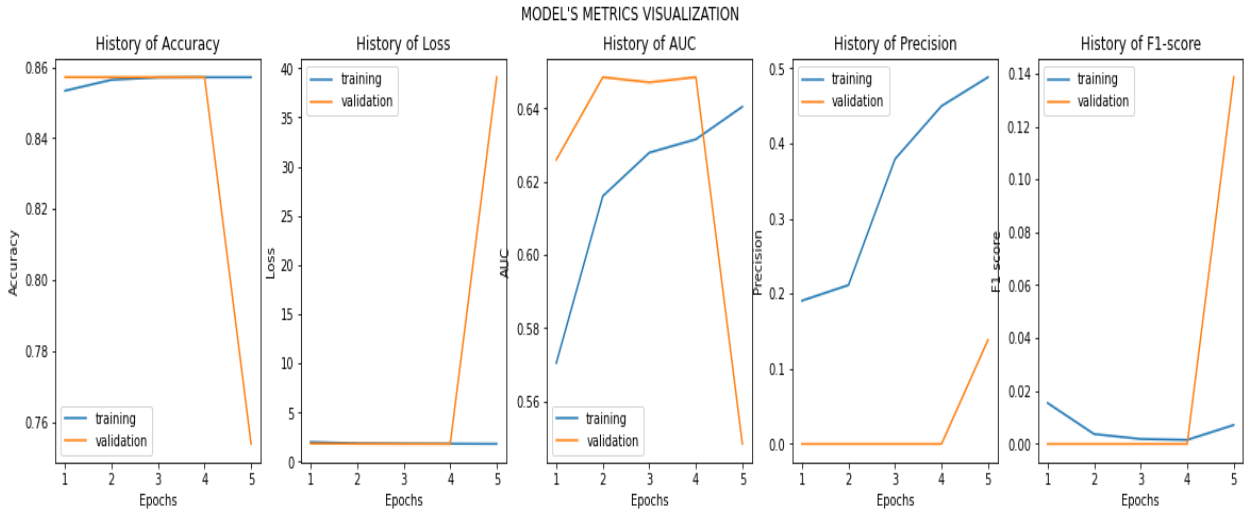
Analytical

We will broadly classify this section into 3 parts.

Image recognition model:

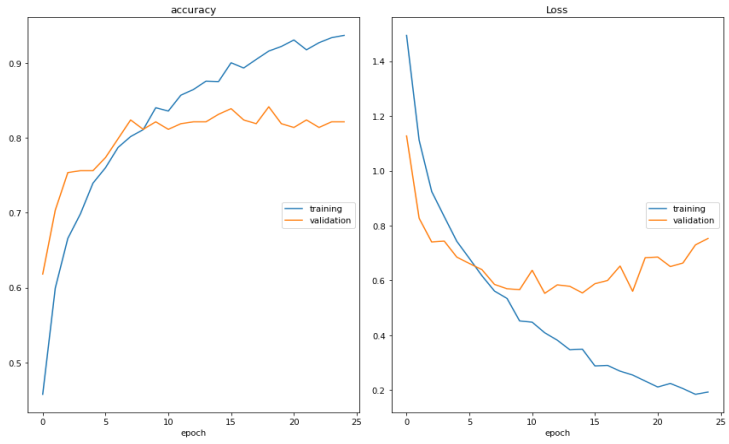
For building the image recognition model, we tried multiple pretrained models. Using the transfer learning technique were successful in achieving substantial accuracy level. In the very beginning of the project, we started with CNN model which gave us the accuracy of 45% on training dataset. Then we switched to AlexNet. This is another predefined model which gave us accuracy of around 84% on training set. We also tried to use MobileNet where the accuracy for training was around 94%. But unfortunately, all the above models gave less than 50% accuracy on testing dataset. At last, we tried another pre-defined model, VGG16. It is a convolution neural net (CNN) architecture which was used to win ILSVR(Imagenet) competition in 2014. With VGG16, our training accuracy was around 95%, while the

testing accuracy was around 82%. We have used different metrics to evaluate the working of our model. The same has been depicted below.



Audio recognition model:

For building the audio recognition model, we started with LSTM. The accuracy for the training model was around 34% while the accuracy for testing was around 28%. We tried adjusting hyper parameters to increase the accuracy by changing the LR and dropout rate. This increased the accuracy rates for both training and testing by 4-5%. Hence, we switched to different model. We build the new model using CNN. Here, we got the training accuracy of about 87% and testing accuracy was about 79%. Here also tuned the hyper parameters to increase the accuracy level. The training accuracy we got now was around 90% and testing accuracy was around 82%. Below are the metrics we used to evaluate our model.



API:

After building these 2 models, now we needed something that can be used to deploy this model on the website. Here comes API in action. The API will take the uploaded video and break it down into image

frames. The frequency for frames is 3 seconds which means the image will be captured from the video after every 3 seconds. After an image is captured, the API will send this image to our image recognition model for emotion prediction. After predicting the emotion of the image, it would save into an array. Then the next image would be captured, and this cycle will continue until the API gets to the end of the video. Once the processing on video is finished, the API will automatically switch to another function where now it will start prediction emotions from the audio. Similar process will be seen here, where the audio is first divided into chunks of 3 seconds. Then these chunks will be sent to the audio model for emotion prediction and the output of this model will be saved into an array. At the end of the audio, the saved arrays for image and audio will be used to produce a graph which will show the count of each emotion on the webpage. The API will also generate a slider which will help the user to move the video/audio to specific moment and check the emotion at that specific time.

RESULTS:

For Image classifier model:

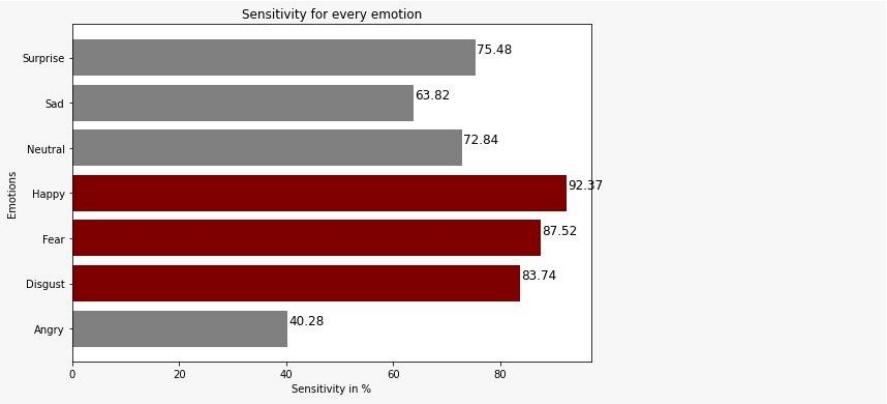
Models tested	Accuracy	Precision	Recall	F1
MobileNet	67.01	69.01	65.29	67.09
AlexNet	52.91	26.61	45	33.44
VGG16	61.43	55.64	48.36	51.74
CNN	43.91	34.46	29.04	31.51
VGG16 using ImageNet	**76.34**	**73.53**	**71.35**	**72.54**

Table 5.1

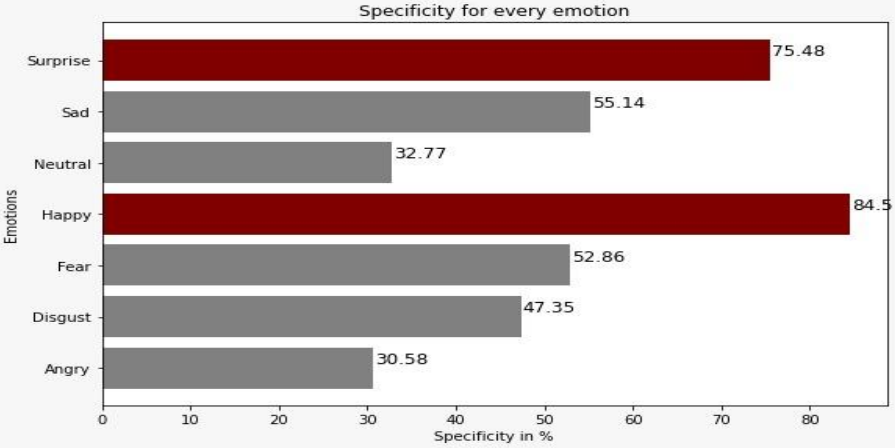
For Image classifier model:

Models tested	Accuracy	Precision	Recall	F1
LSTM	45	30.96	44	36.34
CNN	**81.57**	**81.91**	**82.51**	**82.2**

Table 5.2

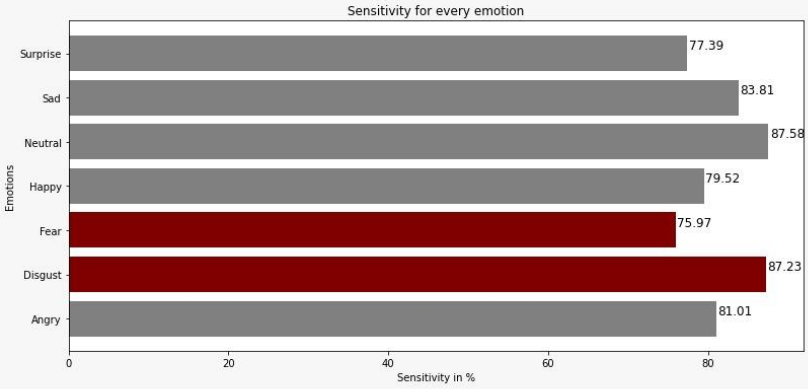


From the plot we can easily calculate sensitivity by dividing correctly predicted true values/all truely pedicted positive values and falsely predicted negative values . In this case, Happy Fear and Disgust emotions are giving us higher sensitivity of 92.37%, 87.52% and 83.74% respectively.



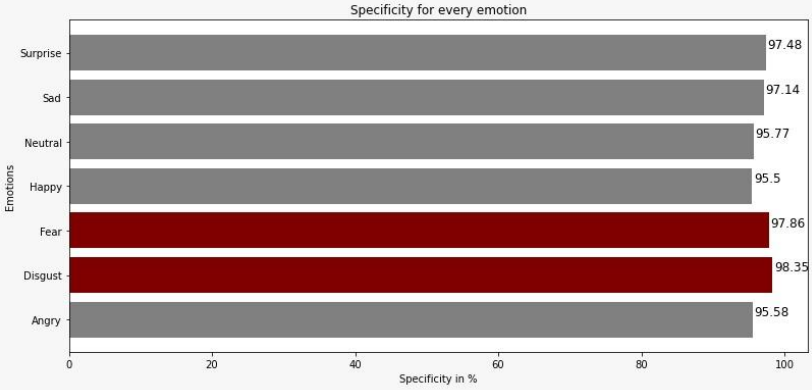
From the plot we can easily calculate sensitivity by dividing Correctly Predicted True Values/all positive results times 100. In this case, Happy and Surprise emotions are giving us higher specificity of 84.50% and 75.48% respectively.

1
2



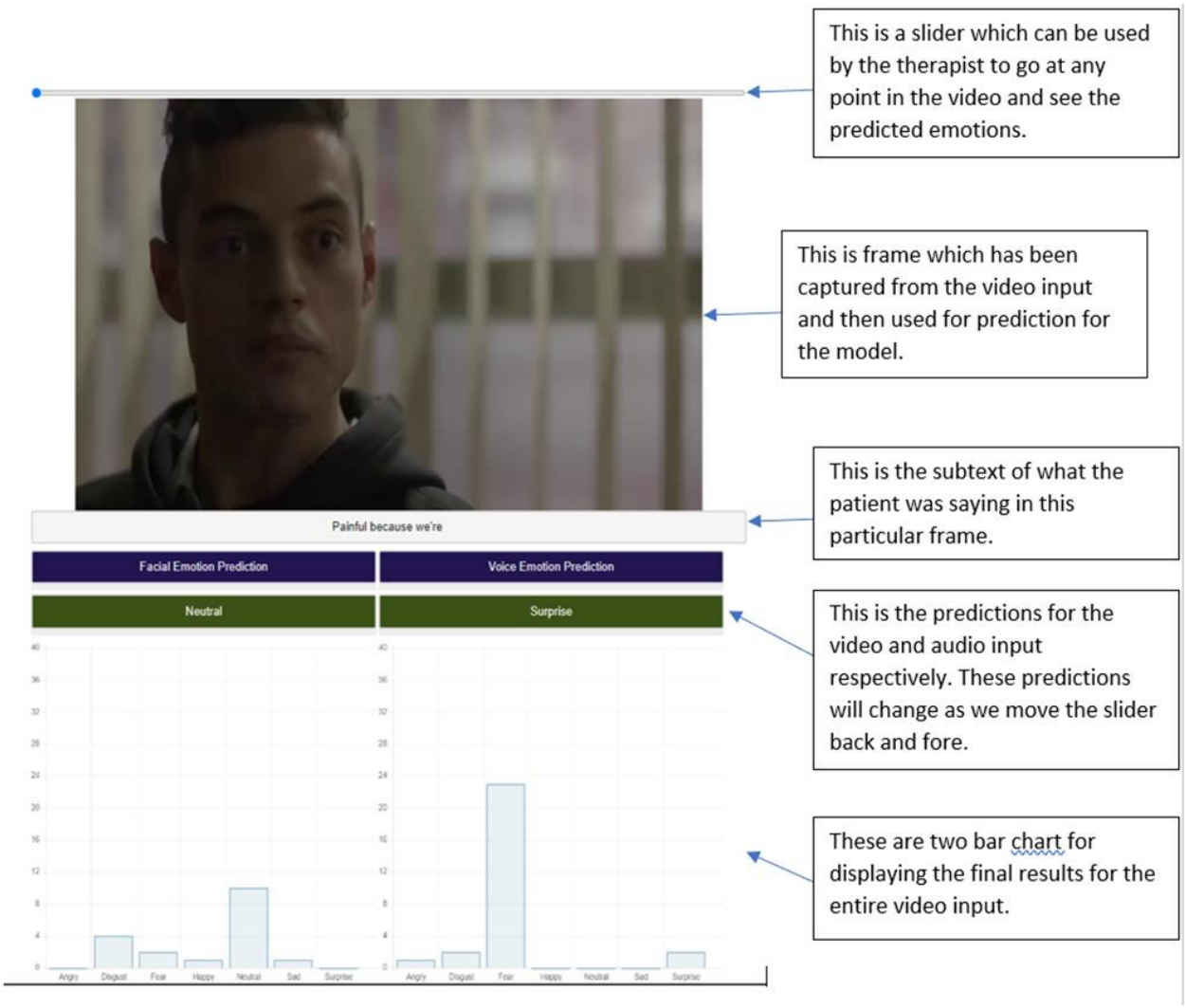
From the plot we can easily calculate Sensitivity by dividing
correctly predicted true positive values/all truely pedicted positive values and predicted flase negative values .
In this case,Neutral and Disgust emotions are giving us higher
sensitivity of 87.58%, 87.23% respectively.

3
4



From the plot we can easily calculate specificity by dividing
correctly predicted true negative values/all truely pedicted negative values and predicted flase positive values .
In this case,Disgust and fear emotions are giving us higher
specificity of 98.35%, 97.86% respectively.

5



DISCUSSION: -

As we see in Table 5.1 and Table 5.2, we can compare results of different models with the best model which we finalized (highlighted).

In image classifier model, the accuracy is about 76.34% and for the audio classification model, the accuracy is about 81.57%. It is simply a ratio of correctly predicted observation to the total observations. The precision for image model and audio model is 73.53% and 81.91% respectively. It is the ratio of correctly predicted positive observations to the total predicted positive observations. The recall value for image model and audio model is 71.35% and 82.51% respectively. Recall is the ratio of correctly predicted positive observations to all observations in actual class. And finally, the F1 score for image model and audio model is 72.51% and 82.20% respectively, which is harmonic mean of precision and recall.

In image classification model, the sensitivity for Happy, Fear and Disgust are the highest when compared to other emotions. In audio classification model, the sensitivity for Fear and Disgust emotion is more when compare to other emotions. This means that these emotions have a high probability of getting predicted correctly. In the same way, in specificity for image classification emotions Happy and

Surprise have a high specificity and in the audio model, Fear and Disgust have a high specificity when compared to other emotion. This means that these emotions have a high proportion of predicting false values.

We have created API in python, using flask, in which we have used our saved models to perform predictions on the video submitted by the therapist. The API will first divide the video into frames and perform predictions on those frames and for the audio part it will firstly, extract the audio from the video input them break the audio file into chunks and then perform predictions on the same. After the prediction is completed, it will be same in an array and dictionary and then send those array and dictionary to the webpage. Webpage will then use these variables passed by the API to display the output in the form of charts and predictions.

CONCLUSIONS:

We were successful in building a model which can help the therapist to recognise the emotion of the client and see the results in the graphical form. All the steps right from uploading the video from the user's end to getting the desired output are preformed without any glitch.

REFERENCES:

1. Jojo George, Kavita Koradiya, Kinjal Bhatt, Krishna Dilip, Tarunkumar Reddy. (2022). Emotion Classification Using Video and Audio Data: CNNs Provides Detailed Analysis with Combination of Data.
2. Al-Darraji, Salah & Berns, Karsten & Rodić, Aleksandar. (2017). Action Unit Based Facial Expression Recognition Using Deep Learning. 540. 413-420. 10.1007/978-3-319-49058-8_45.
3. Bagus Tris Atmaja, Kiyoaki Shirai and Masato Akagi. (2020). Speech Emotion Recognition Using Speech Feature and Word Embedding.